(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **2002182682 A**

(43) Date of publication of application: **26.06.02**

(51) Int. Cl **G10L 15/06**
G10L 15/14
G10L 21/04
G10L 13/06
G10L 15/02

(21) Application number: **2000382371**

(22) Date of filing: **15.12.00**

(71) Applicant: **SHARP CORP**

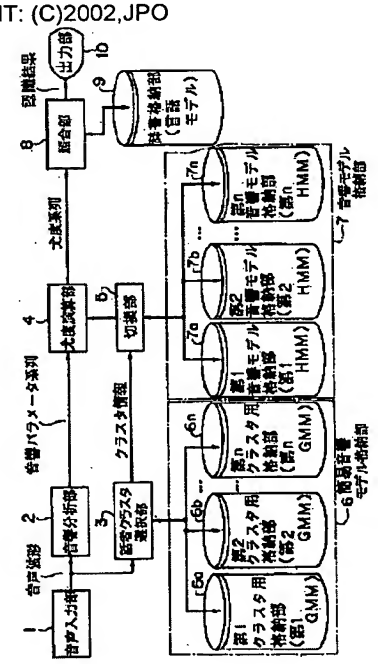
(72) Inventor: **YAMAGUCHI KOICHI**
HACHIMAN YOICHIRO

(54) **SPEAKER CHARACTERISTIC EXTRACTOR, SPEAKER CHARACTERISTIC EXTRACTION METHOD, SPEECH RECOGNIZER, SPEECH SYNTHESIZER AS WELL AS PROGRAM RECORDING MEDIUM**

(57) Abstract:

PROBLEM TO BE SOLVED: To extract speaker characteristics with good accuracy from a smaller quantity of utterance data.

SOLUTION: Acoustic models are stored in a first acoustic model storage section 7a to an n-th acoustic model storage section 7n by each of n pieces of speaker clusters in the acoustic model storage sections 7. The vocal tract length normalization coefficient α determined by estimating likelihood by equation (a) according to a reference of maximizing the likelihood of the acoustic models of learning speakers for the acoustic models of all the learning speakers by using a nonlinear frequency warping obtained by applying a correction factor β to vocal tract length normalization coefficient α is used for clustering of the learning speakers of this case as the distance between the respective learning speakers. The distances between the respective learning speakers are set in accordance with the information on the vocal tract lengths which are the fluctuating factors of the physiological characteristics and the correction information of the ways and habits of the utterance, by which the learning speakers are clustered with the speaker characteristics extracted with good accuracy by taking the speakers' habits into consideration from a smaller quantify of the utterance data as the distances between the respective learning speakers.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（A）**　(11)特許出願公開番号
特開2002－182682
（P2002－182682A）
(43)公開日　平成14年６月26日(2002.6.26)

(51)Int.Cl.[7]	識別記号	FI		テーマコード（参考）
Ｇ１０Ｌ　15/06		Ｇ１０Ｌ　3/00	５２１Ｓ	５Ｄ０１５
15/14			５３５Ｃ	５Ｄ０４５
21/04		3/02	Ａ	
13/06		5/04	Ｆ	
15/02		9/02	３０１Ａ	

審査請求　未請求　請求項の数13　ＯＬ　（全 18 頁）　最終頁に続く

(21)出願番号　特願2000－382371(P2000－382371)

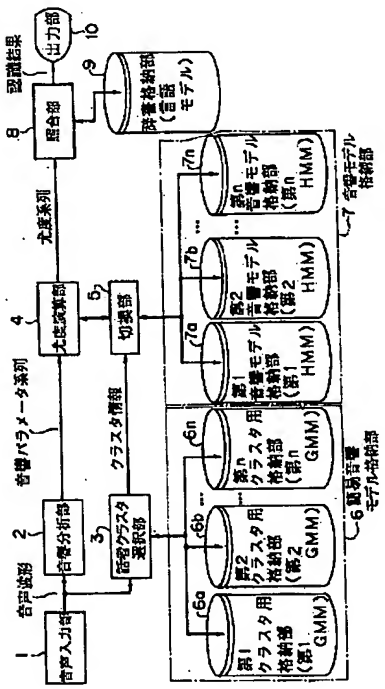
(22)出願日　平成12年12月15日(2000.12.15)

(71)出願人　000005049
シャープ株式会社
大阪府大阪市阿倍野区長池町22番22号
(72)発明者　山口　耕市
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内
(72)発明者　八幡　洋一郎
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内
(74)代理人　100062144
弁理士　青山　葆　（外１名）

Ｆターム(参考)　5D015 FF04 GG04
5D045 AA07

(54)【発明の名称】　話者特徴抽出装置および話者特徴抽出方法、音声認識装置、音声合成装置、並びに、プログラム記録媒体

(57)【要約】
【課題】　より少ない発声データから精度良く話者特徴を抽出する。
【解決手段】　音響モデル格納部７には、ｎ個の話者クラスタ別に第１音響モデル格納部７ａ～第ｎ音響モデル格納部７ｎに音響モデルを格納する。その場合における学習話者のクラスタリングは、各学習話者間の距離として、声道長正規化係数αに補正係数βを適用した非線形周波数ワーピング関数を用いて、全学習話者の音響モデルに対する学習話者の音響モデルの尤度を最大にするという基準に従って、式(１)によって最尤推定して求めた声道長正規化係数αを用いる。こうして、上記各学習話者間の距離を、生理的な特徴の変動要因である声道長の情報と発声の仕方や癖による影響の補正情報とに基づいて設定することによって、より少量の発声データから発話者の癖を考慮して精度良く抽出された話者特徴を各学習話者間の距離として学習話者をクラスタリングできるのである。



1

【特許請求の範囲】

【請求項１】　入力音声から、標準話者の音声のスペクトルに対して上記入力音声のスペクトルの周波数軸を伸縮する際の伸縮係数αを話者特徴として抽出する話者特徴抽出装置において、

音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で上記伸縮係数αに補正係数βを乗じて上記伸縮係数αに対して部分的に補正を行った非線形周波数ワーピング関数を用いて、標準話者の音声パターンに対して入力話者の音声パターンの尤度を最大にするという基準に従って、上記伸縮係数αを求める伸縮係数取得手段を備えたことを特徴とする話者特徴抽出装置。

【請求項２】　請求項１に記載の話者特徴抽出装置において、

上記伸縮係数取得手段は、上記非線形周波数ワーピング関数を用いて、標準話者の音声パターンに対して入力話者の音声パターンの尤度を最大にするという基準に従って、上記補正係数βをも求めるようになっていることを特徴とする話者特徴抽出装置。

【請求項３】　標準話者の音声スペクトルに対して学習話者の音声スペクトルの周波数軸を伸縮する際の伸縮係数αに基づいて上記学習話者をクラスタリングし、各話者クラスタに属する学習話者群の音声パターンに基づいて作成された音響モデルを、各話者クラスタ別の音響モデル格納部に格納した音声認識装置であって、

上記伸縮係数αは、音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で上記伸縮係数αに補正係数βを乗じて上記伸縮係数αに対して部分的に補正を行った非線形周波数ワーピング関数を用いて、上記標準話者の音声パターンに対して学習話者の音声パターンの尤度を最大にするという基準に従って求められていることを特徴とする音声認識装置。

【請求項４】　請求項３に記載の音声認識装置において、

上記学習話者のクラスタリングは、上記伸縮係数αと補正係数βとの２次元平面に対して行われており、

上記補正係数βは、上記非線形周波数ワーピング関数を用いて、上記標準話者の音声パターンに対して学習話者の音声パターンの尤度を最大にするという基準に従って求められていることを特徴とする音声認識装置。

【請求項５】　請求項３あるいは請求項４に記載の音声認識装置において、

上記話者クラスタは、

上記伸縮係数αによる１次元空間、あるいは、上記伸縮係数αと補正係数βとによる２次元空間を、所定のクラスタ数にクラスタリングした初期クラスタと、

上記各初期クラスタの境界を含んで上記各初期クラスタにオーバーラップするオーバーラップクラスタで構成されていることを特徴とする音声認識装置。

2

【請求項６】　標準話者の音声スペクトルに対して入力話者の音声スペクトルの周波数軸を伸縮する際の伸縮係数αを用いて入力話者の音声スペクトルの周波数軸を伸縮することによって上記入力話者の音声を正規化する正規化手段を有する音声認識装置において、

上記正規化手段は、

音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で上記伸縮係数αに補正係数βを乗じて上記伸縮係数αに対して部分的に補正を行った非線形周波数ワーピング関数を用いて、標準話者の音声パターンに対して入力話者の音声パターンの尤度を最大にするという基準に従って、上記伸縮係数αと補正係数βとを推定する周波数ワーピング関数推定手段と、

上記推定された伸縮係数αと補正係数βを係数とする上記非線形周波数ワーピング関数を用いて、上記入力話者の音声スペクトルの周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴とする音声認識装置。

【請求項７】　標準話者の音声スペクトルに対して入力話者の音声スペクトルの周波数軸を伸縮する際の伸縮係数αを用いて音声のスペクトルの周波数軸を伸縮することによって音響モデルを入力話者に話者適応させる話者適応手段を有する音声認識装置において、

上記話者適応手段は、

音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で上記伸縮係数αに補正係数βを乗じて上記伸縮係数αに対して部分的に補正を行った非線形周波数ワーピング関数を用いて、標準話者の音声パターンに対して入力話者の音声パターンの尤度を最大にするという基準に従って、上記伸縮係数αと補正係数βとを推定する周波数ワーピング関数推定手段と、

上記推定された伸縮係数αの逆数と補正係数βの逆数とを係数とする上記非線形周波数ワーピング関数を用いて、上記音響モデルの周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴とする音声認識装置。

【請求項８】　請求項６あるいは請求項７に記載の音声認識装置において、

上記周波数ワーピング関数推定手段は、上記入力話者の音声パターンの代わりに、標準話者の音響モデルを上記入力話者の音声パターンに話者適応させて作成された入力話者用の適応音響モデルを用いるようになっていることを特徴とする音声認識装置。

【請求項９】　請求項４乃至請求項８の何れか一つに記載の音声認識装置において、

上記補正係数βは、音響モデルの状態や音素等のサブワード単位に求められ、上記サブワード毎に決定されていることを特徴とする音声認識装置。

【請求項１０】　入力話者の音声スペクトルに対して標準話者の音声スペクトルの周波数軸を伸縮する際の伸縮

係数αを用いて音声のスペクトルの周波数軸を伸縮することによって、標準話者の音声素片を接続して成る合成音声の声質を発話者の声質に変換する声質変換手段を有する音声合成装置において、

上記声質変換手段は、

音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で上記伸縮係数αに補正係数βを乗じて上記伸縮係数αに対して部分的に補正を行った非線形周波数ワーピング関数を用いて、標準話者の音声パターンに対して上記発話者の音声パターンの尤度を最大にするという基準に従って、上記伸縮係数αと補正係数βとを推定する周波数ワーピング関数推定手段と、

上記推定された伸縮係数αの逆数と補正係数βの逆数とを係数とする上記非線形周波数ワーピング関数を用いて、上記音声素片の周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴とする音声合成装置。

【請求項１１】　請求項１０に記載の音声合成装置において、

上記周波数ワーピング関数推定手段は、上記補正係数βを、音響モデルの状態や音素等のサブワード単位で求め、そのサブワード毎に推定するようになっていることを特徴とする音声合成装置。

【請求項１２】　入力音声から、標準話者の音声のスペクトルに対して上記入力音声のスペクトルの周波数軸を伸縮する際の伸縮係数αを話者特徴として抽出する話者特徴抽出方法において、

音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で上記伸縮係数αに補正係数βを乗じて上記伸縮係数αに対して部分的に補正を行った非線形周波数ワーピング関数を用いて、標準話者の音声パターンに対して入力話者の音声パターンの尤度を最大にするという基準に従って、上記伸縮係数αを求めることを特徴とする話者特徴抽出方法。

【請求項１３】　コンピュータを、

請求項１における上記伸縮係数取得手段として機能させる話者特徴抽出処理プログラムが記録されたことを特徴とするコンピュータ読出し可能なプログラム記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、標準話者の音声スペクトルに対する入力音声スペクトルの周波数軸の線形伸縮係数を話者特徴として抽出する話者特徴抽出装置および話者特徴抽出方法、その抽出方法を用いた音声認識装置，音声合成装置、並びに、話者特徴抽出処理プログラムを記録したプログラム記録媒体に関する。

【０００２】

【従来の技術】従来より、隠れマルコフモデル(Hidden Markov Model：以下、ＨＭＭと言う)を用いた音声認識方法の開発が近年盛んに行われている。このＨＭＭは、

大量の音声データから得られる音声の統計的特徴を確率的にモデル化したものであり、このＨＭＭを用いた音声認識方法の詳細は、中川聖一著「確率モデルによる音声認識」(電子情報通信学会)に詳しい。このＨＭＭに基づく話者適応や話者正規化に関する研究が行われており、特に声道長に基づく話者正規化が盛んに研究されて効果が出ている。声道長の差は話者間の主な変動要因であり、声道長は従来の話者適応法に比べて１個のパラメータあるいは極めて少ないパラメータで音声の特徴を表現できることから、声道長にはより少量の学習データで効率良く正規化できるというメリットがある。

【０００３】ところで、標準話者の音声パターンに対する入力話者の音声サンプルの尤度を最大にするという基準(最尤推定)に従って、上記音声サンプルにおける周波数軸の線形伸縮係数α(声道長正規化係数)を求める。そして、この声道長正規化係数αを用いて入力話者の音声サンプルの周波数軸を線形伸縮することで話者正規化する技術が提案されている(例えば、AT&T Bell Labs. Li Lee, Richard C. Rose, “Speaker Normalization using Efficient Frequency Warping Procedures”, pp. 353-356 ICASSP96 (1996))。また、特開平１１－３２７５９２号公報においては、声道を前室と後室との２つの室に分け、各室に対応した２つの周波数軸線形伸縮係数αを用いて話者正規化する技術が開示されている。

【０００４】尚、上記話者適応は標準となる音響モデルを入力話者に対して適応(つまり正規化)させる技術であり、話者正規化とは表裏一体の関係にある。

【０００５】また、話者クラスタリングを用いた音声認識方法がある。この音声認識方法においては、学習話者間の距離を定義して学習話者をクラスタリングしておき、クラスタ毎にそのクラスタに属する学習話者群の音声データを用いて音響モデルを作成する。そして、認識時には、入力音声に最適なクラスタを選択し、そのクラスタの音響モデルを用いて認識処理を行うのである。その場合における学習話者間の距離として上記声道長の周波数軸線形伸縮係数を用いる音声認識装置が提案されている(特開平１１－１７５０９０号公報)。この公報においては、声道を前室と後室との２つの室に分け、各室に対応した２つの周波数軸線形伸縮係数を用いて学習話者をクラスタリングするようにしている。

【０００６】さらに、音声合成における声質変換に関する従来技術として、音声認識の話者適応技術を用いてスペクトルの写像を行なう方法が提案されている。例えば、ベクトル量子化(ＶＱ)コードブックマッピング法をベースとした話者適応技術を用いる方法(特開平７－１０４７９２号公報)や、ＶＦＳ(Vector Field Smoothing)法をベースとした話者適応技術を用いる方法(橋本誠，樋口宣男：“話者選択と移動ベクトル場平滑化を用いた声質変換のためのスペクトル写像”，信学技報，SP95-1，p.p.1-8，May 1995)等がある。

5

【０００７】

【発明が解決しようとする課題】しかしながら、上記従来の声道長に基づく話者適応や話者正規化には、以下のような問題がある。すなわち、声道長に基づく話者適応や話者正規化は極めて少ないパラメータ数で音声の特徴を表現できるとは言うものの、話者内変動と呼ばれるその時の発話の仕方や癖等の影響を受け易い。したがって、必ずしも少ない学習サンプルから安定して声道長を抽出できるとは限らない。そして、現在一般に用いられている音声データベースには、身長, 出身地, 性別, 年齢等の情報しか付与されていないのである。

【０００８】実際の声道長はＭＲＩ(磁気共鳴画像診断装置)で測定しなければ分からないため、現時点においては直ちに真の声道長を知るのは困難な状況にある。上記特開平１１－３２７５９２号公報および特開平１１－１７５０９０号公報では、声道パラメータを得るために入力音声のフォルマント周波数を用いている。しかしながら、一般的にフォルマント周波数を全自動で求めることは困難であり、上記特開平１１－３２７５９２号公報に開示された線形伸縮係数を用いた話者正規化方法や上記特開平１１－１７５０９０号公報に開示された線形伸縮係数を用いた音声認識装置では、実時間性に欠けるという問題がある。

【０００９】さらに、発声の仕方や一部の発音器官の形状による違いもフォルマント周波数のずれとなって現れる。そのため、少ない発声データから声道長を正規化するのは一般には困難である。また、声道長の伸縮(すなわち周波数軸のワーピング)を線形関数やそれに類似した関数で表現している。そのために、全区間に対して一様に周波数ワープを作用させることになり、声道長の差の影響を受け難い音素や無音部まで正規化(すなわち変形)されてしまうという問題もある。

【００１０】また、上記特開平１１－１７５０９０号公報のごとく、話者クラスタリングを用いた音声認識のアプローチも盛んに試みられているが、大きな性能改善は達成できていない。不特定話者(ＳＩ)音響モデル(すなわち男女共通の音響モデル)をベースラインとすると、男女別(ＧＤ)音響モデルは最もシンプルながら性能向上量が最も大きい。しかしながら、話者クラスタによって更なる細分化(クラスタ化)を行っても効果は薄いという報告がなされており、その場合における単語誤り率(ＷＥＲ: Word Error Rate)の削減は１０％～２０％程度に留まっている。これは、話者間の距離を定義する適当な尺度がないために上手くクラスタリングできなかったり、クラスタを増やすと１つのクラスタ当りの学習話者数が少なくなってロバスト性に欠けたりするためである。

【００１１】さらに、何れの音響モデルの場合も、各話者クラスタの境界領域では学習サンプルが希薄だったり段差ができたりしているため上手く学習されていない。

6

したがって、入力話者が各クラスタの境界付近に位置する場合には、認識率が劣化するという問題(所謂、hard decision問題)が生ずることになる。尚、個々の学習話者の音響モデル間の距離でクラスタリングを行った場合は、クラスタを木構造にし、入力話者が二つのクラスタの境界付近に位置する場合は上記２つのクラスタの上位ノードのクラスタの音響モデルを採用する方法もある。しかしながら、この方法の場合には、二つのクラスタの境界付近に位置する入力話者に対しては上位ノードの音響モデルを使用するためによりブロードな音響モデルとなってしまい、高い認識率は得にくいのである。

【００１２】以上のごとく、上記話者適応(話者正規化)においては少ない発声データから音響モデルを精度良く適応できないため、誤り率を半減させるためには数十単語以上の発声データが必要となり、学習話者に負担を強いることになるという問題がある。また、音声合成における声質変換の場合にも、同様に少ない発声データからは精度良く声質が得られないという問題がある。

【００１３】そこで、この発明の目的は、より少ない発声データから精度良く話者特徴を抽出できる話者特徴抽出装置および話者特徴抽出方法、その抽出方法を用いた音声認識装置, 音声合成装置、並びに、話者特徴抽出処理プログラムを記録したプログラム記録媒体を提供することにある。

【００１４】

【課題を解決するための手段】上記目的を達成するため、第１の発明は、入力音声から, 標準話者の音声のスペクトルに対して上記入力音声のスペクトルの周波数軸を伸縮する際の伸縮係数αを話者特徴として抽出する話者特徴抽出装置において、音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で上記伸縮係数αに補正係数βを乗じて上記伸縮係数αに対して部分的に補正を行った非線形周波数ワーピング関数を用いて、標準話者の音声パターンに対して入力話者の音声パターンの尤度を最大にするという基準に従って、上記伸縮係数αを求める伸縮係数取得手段を備えたことを特徴としている。

【００１５】上記構成によれば、音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で伸縮係数αに補正係数βを乗じて補正を行った非線形周波数ワーピング関数を用いて、最尤推定によって上記伸縮係数α求め、その求められた伸縮係数αをもって話者特徴としている。したがって、生理的な特徴の変動要因である声道長の情報に対して発声の仕方や癖による影響の補正が行われて、より話者に適合した特徴が抽出される。その際に、話者の発声の仕方や癖による影響が予め補正されている。したがって、上記発声の仕方や癖を表す発声データを必要とはせず、より少量の発声データから精度良く話者特徴が抽出される。

【００１６】また、上記第１の発明の話者特徴抽出装置

は、上記伸縮係数取得手段を、上記非線形周波数ワーピング関数を用いて、標準話者の音声パターンに対して入力話者の音声パターンの尤度を最大にするという基準に従って、上記補正係数βをも求めるように成すことが望ましい。

【００１７】上記構成によれば、話者特徴として、上記伸縮係数αに加えて、広母音の第２フォルマントの存在領域以下の低い周波数領域に対する補正係数βもが抽出される。したがって、さらに話者に適合した特徴が抽出される。

【００１８】また、第２の発明は、標準話者の音声スペクトルに対して学習話者の音声スペクトルの周波数軸を伸縮する際の伸縮係数αに基づいて上記学習話者をクラスタリングし，各話者クラスタに属する学習話者群の音声パターンに基づいて作成された音響モデルを各話者クラスタ別の音響モデル格納部に格納した音声認識装置であって、上記伸縮係数αは、音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で上記伸縮係数αに補正係数βを乗じて上記伸縮係数αに対して部分的に補正を行った非線形周波数ワーピング関数を用いて、上記標準話者の音声パターンに対して学習話者の音声パターンの尤度を最大にするという基準に従って求められていることを特徴としている。

【００１９】上記構成によれば、学習話者をクラスタリングする際における各学習話者間の距離として、上記非線形周波数ワーピング関数を用いて最尤推定によって求めた上記伸縮係数αが用いられる。こうして、生理的な特徴の変動要因である声道長の情報に対して発声の仕方や癖による影響の補正が行われて、より学習話者の音声パターンに適合した距離を用いて上記クラスタリングが行われる。その際に、話者の発声の仕方や癖による影響が予め補正されているので、上記発声の仕方や癖を表す発声データを必要とはせず、より少量の発声データから学習話者間の距離が得られるのである。

【００２０】また、上記第２の発明の音声認識装置は、上記学習話者のクラスタリングを，上記伸縮係数αと補正係数βとの２次元平面に対して行うようになっており、上記補正係数βを，上記非線形周波数ワーピング関数を用いて，上記標準話者の音声パターンに対して学習話者の音声パターンの尤度を最大にするという基準に従って求めることが望ましい。

【００２１】上記構成によれば、上記学習話者間の距離として、上記伸縮係数αに加えて、上記補正係数βも用いられる。したがって、さらに話者に適合した距離を用いてクラスタリングが行われる。

【００２２】また、上記第２の発明の音声認識装置は、上記話者クラスタを、上記伸縮係数αによる１次元空間あるいは上記伸縮係数αと補正係数βとによる２次元空間を，所定のクラスタ数にクラスタリングした初期クラスタと、上記各初期クラスタの境界を含んで上記各初期クラスタにオーバーラップするオーバーラップクラスタとで構成することが望ましい。

【００２３】上記構成によれば、話者クラスタは、所定数の初期クラスタと上記各初期クラスタにオーバーラップするオーバーラップクラスタとで構成されている。したがって、学習サンプルが希薄だったり段差ができ易い上記各初期クラスタの境界領域は、何れかのオーバーラップクラスタに含まれることになり、上記各初期クラスタの境界領域において認識率が劣化するという「hard decision問題」が解消される。

【００２４】また、第３の発明は、標準話者の音声スペクトルに対して入力話者の音声スペクトルの周波数軸を伸縮する際の伸縮係数αを用いて入力話者の音声スペクトルの周波数軸を伸縮することによって上記入力話者の音声を正規化する正規化手段を有する音声認識装置において、上記正規化手段は、音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で上記伸縮係数αに補正係数βを乗じて上記伸縮係数αに対して部分的に補正を行った非線形周波数ワーピング関数を用いて，標準話者の音声パターンに対して入力話者の音声パターンの尤度を最大にするという基準に従って，上記伸縮係数αと補正係数βとを推定する周波数ワーピング関数推定手段と、上記推定された伸縮係数αと補正係数βとを係数とする上記非線形周波数ワーピング関数を用いて，上記入力話者の音声スペクトルの周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴としている。

【００２５】上記構成によれば、周波数ワーピング関数推定手段によって、上記非線形周波数ワーピング関数を用いて、上記伸縮係数αと補正係数βとが最尤推定される。そして、周波数ワープ手段によって、推定されたαとβとを係数とする上記非線形周波数ワーピング関数を用いて入力話者が正規化される。こうして、生理的な特徴の変動要因である声道長の情報に対して発声の仕方や癖による影響の補正が行われて、より標準話者の音声スペクトルに近づくように話者の正規化が行われる。その際に、話者の発声の仕方や癖による影響が予め補正されているので、上記発声の仕方や癖を表す発声データを必要とはせず、より少量の発声データに基づいて話者正規化が行われるのである。

【００２６】また、第４の発明は、標準話者の音声スペクトルに対して入力話者の音声スペクトルの周波数軸を伸縮する際の伸縮係数αを用いて音声のスペクトルの周波数軸を伸縮することによって音響モデルを入力話者に話者適応させる話者適応手段を有する音声認識装置において、上記話者適応手段は、音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で上記伸縮係数αに補正係数βを乗じて上記伸縮係数αに対して部分的に補正を行った非線形周波数ワーピング関数を用いて，標準話者の音声パターンに対して入

9

力話者の音声パターンの尤度を最大にするという基準に従って，上記伸縮係数αと補正係数βとを推定する周波数ワーピング関数推定手段と、上記推定された伸縮係数αの逆数と補正係数βの逆数とを係数とする上記非線形周波数ワーピング関数を用いて，上記音響モデルの周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴としている。

【００２７】上記構成によれば、周波数ワーピング関数推定手段によって、上記非線形周波数ワーピング関数を用いて、上記伸縮係数αと補正係数βとが最尤推定される。そして、周波数ワープ手段によって、この推定された伸縮係数αの逆数と補正係数βの逆数とを係数とする上記非線形周波数ワーピング関数を用いて、音響モデルが入力話者に話者適応される。こうして、生理的な特徴の変動要因である声道長の情報に対して発声の仕方や癖による影響の補正が行われて、より入力話者の音声スペクトルに近づくように話者適応が行われる。その際に、話者の発声の仕方や癖による影響が予め補正されているので、上記発声の仕方や癖を表す発声データを必要とはせず、より少量の発声データに基づいて話者適応が行われるのである。

【００２８】また、上記第３の発明あるいは第４の発明の音声認識装置は、上記周波数ワーピング関数推定手段を、上記入力話者の音声パターンの代わりに、標準話者の音響モデルを上記入力話者の音声パターンに話者適応させて作成された入力話者用の適応音響モデルを用いるように成すことが望ましい。

【００２９】上記構成によれば、上記伸縮係数αと補正係数βとの推定に際して、入力話者の音声パターンそのものではなく、標準話者の音響モデルを入力話者の音声パターンに話者適応させた適応音響モデルを用いるので、入力話者の音声パターン数が少ない場合でも対処可能となる。さらに、上記適応音響モデルの状態毎に補正係数βを制御して、話者の発声の仕方や癖による入力音声パターンのずれを木目細かく補正することが可能になる。

【００３０】また、上記第２の発明乃至第４の発明の何れか一つの発明の音声認識装置は、上記補正係数βを、音響モデルの状態や音素等のサブワード単位に求め、上記サブワード毎に決定することが望ましい。

【００３１】上記構成によれば、上記補正係数βがサブワード単位に変更されて、話者の発声の仕方や癖による入力音声パターンのずれが木目細かく補正される。

【００３２】また、第５の発明は、標準話者の音声スペクトルに対して入力話者の音声スペクトルの周波数軸を伸縮する際の伸縮係数αを用いて音声のスペクトルの周波数軸を伸縮することによって，標準話者の音声素片が接続されて成る合成音声の声質を発話者の声質に変換する声質変換手段を有する音声合成装置において、上記声質変換手段は、音声のスペクトルにおける広母音の第２

10

フォルマントの存在領域以下の低い周波数領域で上記伸縮係数αに補正係数βを乗じて上記伸縮係数αに対して部分的に補正を行った非線形周波数ワーピング関数を用いて，標準話者の音声パターンに対して上記発話者の音声パターンの尤度を最大にするという基準に従って，上記伸縮係数αと補正係数βとを推定する周波数ワーピング関数推定手段と、上記推定された伸縮係数αの逆数と補正係数βの逆数とを係数とする上記非線形周波数ワーピング関数を用いて，上記音声素片の周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴としている。

【００３３】上記構成によれば、周波数ワーピング関数推定手段によって、上記非線形周波数ワーピング関数を用いて、最尤推定によって上記伸縮係数αと補正係数βとが推定される。そして、周波数ワープ手段によって、推定された伸縮係数αの逆数と補正係数βの逆数とを係数とする上記非線形周波数ワーピング関数を用いて、標準話者の音声素片の周波数軸が伸縮される。こうして、生理的な特徴の変動要因である声道長の情報に対して発声の仕方や癖による影響の補正が行われて、上記合成音声の声質が発話者の声質に変換される。その際に、話者の発声の仕方や癖による影響が予め補正されているので、上記発声の仕方や癖を表す発声データを必要とはせず、より少量の発声データに基づいて声質変換が行われる。

【００３４】また、上記第５の発明の音声合成装置は、上記周波数ワーピング関数推定手段を、音響モデルの状態や音素等のサブワード単位に上記補正係数βを求め、上記サブワード毎に上記補正係数βを推定するように成すことが望ましい。

【００３５】上記構成によれば、上記補正係数βがサブワード単位に変更されて、発話者の発声の仕方や癖による入力音声パターンのずれが木目細かく補正される。

【００３６】また、第６の発明は、入力音声から，標準話者の音声のスペクトルに対して上記入力音声のスペクトルの周波数軸を伸縮する際の伸縮係数αを話者特徴として抽出する話者特徴抽出方法において、音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で上記伸縮係数αに補正係数βを乗じて上記伸縮係数αに対して部分的に補正を行った非線形周波数ワーピング関数を用いて、標準話者の音声パターンに対して入力話者の音声パターンの尤度を最大にするという基準に従って、上記伸縮係数αを求めることを特徴としている。

【００３７】上記構成によれば、上記非線形周波数ワーピング関数を用いて、最尤推定によって伸縮係数αが話者特徴として求められる。したがって、生理的な特徴の変動要因である声道長の情報に対して発声の仕方や癖による影響の補正が行われて、より話者に適合した特徴が抽出されるのである。その際に、話者の発声の仕方や癖

による影響が予め補正されている。したがって、より少量の発声データから良質の話者特徴が抽出される。

【００３８】また、第７の発明のプログラム記録媒体は、コンピュータを、上記第１の発明の伸縮係数取得手段として機能させる話者特徴抽出処理プログラムが記録されていることを特徴としている。

【００３９】上記構成によれば、上記第１の発明の場合と同様に、生理的な特徴の変動要因である声道長の情報に対して発声の仕方や癖による影響の補正が行われて、より話者に適合した特徴が抽出される。その際に、話者の発声の仕方や癖による影響が予め補正されているため、より少量の発声データから良質の話者特徴が抽出される。

【００４０】

【発明の実施の形態】以下、この発明を図示の実施の形態により詳細に説明する。

＜第１実施の形態＞図１は、本実施の形態の音声認識装置におけるブロック図である。尚、この音声認識装置は、話者クラスタリング方式を用いた音声認識装置である。音声入力部１において、マイクから入力された音声はディジタル波形に変換されて音響分析部２および話者クラスタ選択部３に入力される。音響分析部２は、入力されたディジタル波形を短い時間間隔(フレーム)毎に周波数分析し、スペクトルを表す音響パラメータのベクトル系列に変換する。ここで、上記周波数分析には、ＭＦＣＣ(メル周波数ＦＦＴ(高速フーリエ変換)ケプストラム)やＬＰＣ(線形予測分析)メルケプストラム等のスペクトルを効率よく表現できる方法が用いられる。こうして得られた音響パラメータ系列は、尤度(音韻類似度)演算部４に送出される。

【００４１】上記話者クラスタ選択部３は、後に詳述する簡易音響モデル格納部６に話者クラスタ別に格納された簡易音響モデル(本実施の形態ではＧＭＭ(ガウシアン混合モデル)を使用)を用いて入力音声の話者クラスタを判定し、クラスタ情報を出力する。切換部５は、後に詳述する音響モデル格納部７に話者クラスタ別に格納された音響モデル(本実施の形態ではHMMを使用)の中から、話者クラスタ選択部３からのクラスタ情報に適合する話者クラスタの音響モデルを切り換え選択して尤度演算部４に送出する。そうすると、尤度演算部４は、音響分析部２からの入力音声の音響パラメータベクトルに対して切換部５からの音響モデルを作用させて、各音韻の状態毎に尤度を算出する。そして、得られた尤度系列を照合部８に送出する。

【００４２】上記照合部８は、上記尤度演算部４からの尤度系列に対して、辞書格納部９に登録された総ての言語モデル(単語)との照合を行ない、各単語のスコアを算出する。そして、上位のスコアを呈する単語を認識候補(認識結果)として出力部１０から出力するのである。

【００４３】ここで、本実施の形態における音響モデル格納部７は、学習話者のクラスタ数ｎに応じて、第１音響モデル格納部７a, 第２音響モデル格納部７b, …, 第ｎ音響モデル格納部７nのｎ個の音響モデル格納部で構成されている。ここで、各音響モデル格納部７a～７nに格納される各音響モデルは、混合ガウス分布型のHMMである。この発明においては、生理的な特徴の変動に対処可能にすることを目的としており、話者性の大局的な安定要因である声道長をクラスタリング対象にするのである。尚、生理的な特徴の変動要因としては、上記声道長以外にも鼻腔, 副鼻腔, 声帯等の多くの要因があり、それらが絡み合って複雑な特徴を成している。したがって、個々の要因を数理的に扱うのは得策ではない。そこで、本実施の形態においては、各要因の複雑な特徴を混合ガウス分布型HMM(音響モデル)で表現するのである。

【００４４】以下、上記音響モデル格納部７に格納される音響モデル群の作成方法について説明する。本実施の形態における音響モデル作成方法は、下記の(１)～(５)の５段階で構成される。

【００４５】(１) 全話者モデルの作成

全学習話者の音声データを用いて学習を行い、混合数が１の音響モデル(全話者モデルと言う)を作成する。ここで、上記学習話者の集合を求めるに際して、男女別に２つの集合に分けてもよい。その場合には、大きく分けて男性用話者クラスタと女性用話者クラスタとの合計２種類の音響モデル群が生成されることになる。尚、上述のような全話者モデルに対して、後に第２実施の形態で述べる話者正規化方式による音声認識時において尤度演算部で用いる混合数が多数の音響モデルを、不特定話者モデルと呼ぶことにする。

【００４６】(２) 特定話者モデルの作成

各学習話者毎に混合数が１の音響モデル(特定話者モデルと言う)を作成する。ここで、各学習話者に関しては、ある程度の量の音声データが整備されているものとする。ここで、１話者当りの音声データ量が多い場合には、HMMの学習アルゴリズムを用いて上記特定話者モデルを作成する。一方、１話者当りの音声データ量が少ない場合は、上記全話者モデルを基に上記ＶＦＳやＭＬＬＲ(MaximumLikelihood Linear Regression)等の手法を用いて話者適応することによって上記特定話者モデルを作成する。

【００４７】(３) 声道長正規化係数αの導出

クラスタリングの基準となる声道長正規化係数αを、補正係数βと共に、各学習話者の特定話者モデルに関して、次に述べる非線形周波数ワーピング関数f()を用いて、式(１)に従って全話者モデルに対する尤度を最大にするという基準で求める。こうして、上記全話者モデルと特定話者モデルとの間の写像関係が、非線形周波数ワーピング関数f()を用いて求めることができるのである。

【００４８】非線形周波数ワーピング関数f()：

13

・ｘ≦θでは、　　f(x)＝αβｘ
（０．８８＜α＜１．１３、０．８＜β≦１）（θ＝１．５kHz～１．８kHz）
・θ＜ｘ≦min(ω/α，ω)では、f(x)＝αｘ　（ω≒４kHz）
・min(ω/α，ω)＜ｘでは、
α＞１のとき　f(x)→(ω/α，ω)と(fs/２，fs/２)とを結ぶ直線
α≦１のとき　f(x)→(ω，αω)と(fs/２，fs/２)とを結ぶ直線
ここで、θ：広母音(日本語の場合「ア」や「オ」に相当)の第２フォルマントが存在する領域における上限周波数
fs：サンプリング周波数

【００４９】ここで、上記非線形周波数ワーピング関数f()における不連続性を考慮して、θ付近において直線「f(x)＝αβｘ」と直線「f(x)＝αｘ」とを接続させる。すなわち、(ｘ，f(x))座標上における(ｋθ，ｋαβθ)と(θ，αθ)とを直線で結ぶのである。ｋは直線f(x)＝αβｘ側の折れ線の頂点を表す定数であり、「０．７」程度の値とする。ここで、上記θは、男性の場合は１．５kHz程度に設定され、女性の場合は１．８kHz程度に設定される。尚、上記θは、話者の広母音における第２フォルマント周波数に応じて話者毎に設定を変えてもよい。また サンプリング周波数fsは、本実施の形態においては８kHz以上を仮定している。すなわち、fs＝１２kHzの場合には、(fs/２，fs/２)は(６kHz，６kHz)となるのである。

【００５０】α＞１である場合における上述のような折れ線で表される非線形周波数ワーピング関数f(x)を図２に示す。αとβとの変動範囲「０．８８＜α＜１．１３」、＊「０．８＜β≦１」は、夫々声道長の分布と声門上部の狭めにより生じるフォルマントの上昇度合いの観測結果に基づいて定めている。発声の仕方や癖によって声門上部の狭めが生じると、広母音の第２フォルマントの存在領域以下の低域のフォルマント周波数が上昇する。そのために、声道長正規化係数αのみでは正しい声道長に写像できないことがある。上記係数βは、この低域のフォルマント周波数の上昇に対する補正項なのである。

14

【００５１】上記音響分析部２による音響分析で得られる音響パラメータや上記音響モデルの出力確率密度関数の引数は、通常ＭＦＣＣやＬＰＣケプストラムである。これらの音響パラメータの各次元はケプストラムと呼ばれる物理量であって、周波数ではない。そこで、上記周波数ワープ処理を行なう際には、学習データである音響パラメータからスペクトルへの変換C^{-1}(ケプストラムの場合は逆cos変換)を行なって周波数次元に変換する。そして、周波数ワープ処理終了後は、逆変換Ｃ(ケプストラムの場合はcos変換)を行なって元の音響パラメータ次元に戻すのである。

【００５２】ここで、標準モデルと入力モデルとの２つの音響モデルの対応する状態間の尤度を、標準モデルの出力確率密度関数ｒｉ()に、入力モデルの出力確率密度関数ｂｉ()の平均値ベクトルμｉを非線形周波数ワーピング関数f()で周波数ワープして得られたベクトルを代入したときの値と定義する。上記各出力確率密度関数は多次元ガウス分布であって、平均値ベクトルと分散ベクトルから成っている。

【００５３】以上のことから、上記声道長正規化係数αは式(１)によって最尤推定できることになる。

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{i \in \Omega} r_i(\mu_i^f) \qquad (0.8 < \beta \leqq 1) \qquad \cdots(1)$$

ここで、　Ω：評価対象の出力確率密度関数集合のインデックス
ri()：標準モデルの第ｉ番目の出力確率密度関数
f()：(α，β)を係数とする非線形周波数ワーピング関数
μ_i^f：入力モデルにおける第ｉ番目の出力確率密度関数ｂｉ()の平均値ベクトルμｉをf()で周波数ワープしたベクトル。すなわち、C(f(C^{-1}(μｉ)))となる。
C^{-1}，Ｃ：音響パラメータからスペクトルへの変換とその逆変換

【００５４】クラスタリングは、上記声道長正規化係数αのみの１次元空間でのクラスタリングと、声道長正規化係数αと補正係数βとの２次元空間でのクラスタリングとの２通りがある。声道長正規化係数αは話者毎に固定されるが、補正係数βは話者内で固定される場合と発話間で固定(＝話者内で変動)される場合の２通りがある。どの範囲で補正係数βを固定するかは話者の発声の癖に依存するため、話者に応じて使い分けるものとする。本実施の形態においては、音響モデルの各状態毎に固定するものとする。

【００５５】(４) 学習話者のクラスタリング
上記声道長正規化係数αを用いて学習話者を所望の数ｎにクラスタリングし、各学習話者の夫々が何れの話者クラスタに属するかを決める。ここで、上記クラスタリングの方法は種々提案されているが、声道長正規化係数αのみでクラスタリングする場合は、１次元空間であるからα軸をｎ個に等分割すればよい。尚、総学習話者数が少ない場合には、各話者クラスタに属する話者数が等しくなるようにｎ個に分割してもよい。声道長正規化係数αと補正係数βとの２次元空間上でクラスタリングする場合は、学習話者をｋ－means法等の手法によってクラスタリングすればよい。

【００５６】ところで、何れの音響モデルの場合も、各話者クラスタの境界領域では学習サンプルが希薄だったり段差ができたりしているために上手く学習されていない。したがって、入力話者が各話者クラスタの境界付近に位置する場合には、認識率が劣化するという「hard de

cision問題」が生じる。そこで、本実施の形態においては、この「hard decision問題」の対策として、学習話者を単純に分割するだけではなくオーバーラップさせて分割するのである。このオーバーラップは声道長に対応しているので物理的にも意味がある。すなわち、先ずn0個の話者クラスタに初期分割した後、n0個の話者クラスタの各境界を中心として初期分割された話者クラスタにオーバーラップする(n0－１)個の話者クラスタに分割するのである。したがって、話者クラスタの数は合計(２n0－１)個となる。図３に、初期分割数n0が「５」の場合のクラスタリング例を示す。縦軸は学習話者の頻度であり、横軸は声道長正規化係数αである。５個の初期分割クラスタの境界を埋めるオーバーラップクラスタの数は４個であるから、総話者クラスタ数は合計９個となる。

【００５７】上記オーバーラップさせるクラスタリングにおいて、オーバーラップのさせ方として、初期分割数n0の異なる話者クラスタを併用してもよい。さらに、分割なしの全話者クラスタや男女別話者クラスタを併用してもよい。例えば、初期分割数n0＝７の話者クラスタに、初期分割数n0＝５の話者クラスタと男女別話者クラスタとを併用すると、(７＋６)＋(５＋４)＋２の合計２４個の話者クラスタとなる。

【００５８】(５) 話者クラスタ別に音響モデルを作成
上記ｎ個の話者クラスタに属する総ての学習話者の音声データを用いて学習を行い、話者クラスタ毎に混合ガウス分布型HMMの音響モデルを作成する。話者クラスタへの初期分割数をn0個とすると、上記オーバーラップクラスタリングによって合計ｎ＝(２n0－１)個の音響モデルが生成されることになる。こうして作成されたｎ個の音響モデルの夫々が、音響モデル格納部７を構成するｎ個の音響モデル格納部７a～７nの何れかに格納されるのである。尚、音響モデル格納部７a～７nの夫々に格納される「１個の音響モデル」とは、文字通り１つの音素の音響モデルを意味するのではなく、全音素に関する音響モデルの総称であることは言うまでも無い。

【００５９】次に、上記簡易音響モデル格納部６について説明する。簡易音響モデル格納部６は、話者のクラスタ数ｎに応じて、第１クラスタ用格納部６a, 第２クラスタ用格納部６b, …, 第ｎクラスタ用格納部６nのｎ個の簡易音響モデル格納部で構成されている。ここで、各クラスタ用格納部６a～６nに格納される各簡易音響モデルはGMMである。尚、GMMは、全音素を１状態で表す多混合連続分布型音響モデルである。

【００６０】そして、上記話者クラスタ選択部３は音響分析手段を内蔵しており、入力音声から抽出された音響パラメータ系列に対して各クラスタ用格納部６a～６nに格納された総てのGMMを作用させて各GMM毎の尤度を算出する。そして、最も大きい尤度を呈するGMMが格納されたクラスタ用格納部６a～６nを表すクラスタ情報を出力するのである。その場合、入力音声の正解音素列をユーザが教える必要がなく、教師なしで話者クラスタを選択することができる。すなわち、エンロールモードがないシステムにおいて有効なのである。

【００６１】ここで、上記話者クラスタ選択の方法には、以下の[ａ]～[ｃ]に示す３通りの方法がある。本実施の形態においては[ｂ]の方法を用いている。

[ａ] 話者クラスタ音響モデル自身の利用

[ｂ] 簡易型音響モデルの利用

[ｃ] 声道長正規化係数αおよび補正係数βの直接推定

【００６２】上記[ａ]の方法は、上記話者クラスタの音響モデル自身の尤度を用いる方法である。入力音声に対して教師語彙が与えられ、各話者クラスタにおける教師語彙の音響モデルを用いて認識処理を行い、各話者クラスタ毎の尤度を算出する。そして、最も大きい尤度を呈する話者クラスタを選択するのである。この選択方法は、エンロールモードにおいて入力音声の正解音素列をユーザが教えるという教師あり選択を基本としている。認識処理と同じ高精度な音響モデルを用いるので計算量は多くなるがエンロールによって正確なクラスタ選択が可能となる。

【００６３】また、上記[ｃ]の方法は、上述した音響モデルの作成方法における(３)の声道長正規化係数αの導出で説明した手法と同様の手法を用いる。但し、特定話者音響モデルからではなく入力音声データから直接求めることになる。つまり、入力音声データに非線形周波数ワーピング関数f()を作用させて、全話者モデルを用いて最尤推定する方法で声道長正規化係数αと補正係数βとを求めるのである。この選択方法は、[ａ]や[ｂ]の選択方法に比して不安定ではあるが、エンロールが可能であり、入力音声サンプルが多量にある場合には有効である。これは、後に第２実施の形態において説明する話者正規化で用いる手法と同じである。

【００６４】上記構成において、入力音声の認識時には以下のように動作する。先ず、話者クラスタ選択部３によって、上述のようにして最適な話者クラスタが選択され、クラスタ情報が切換部５に送出される。次に、尤度演算部４によって、切換部５で切り換え選択された話者クラスタの音響モデルを用いて尤度演算が行われ、得られた尤度系列が照合部８に送出される。そして、照合部８によって、ビタビサーチ等の探索アルゴリズムが用いられて辞書格納部９の言語モデルとの照合が行われ、各単語のスコアが算出される。尚、本実施の形態においては、照合部８による照合処理の前段処理が訴求点であるから、照合部８に関する詳細な説明は省略する。

【００６５】上述のように、本実施の形態においては、上記音響モデル格納部７に格納する音響モデル群の作成に当って、学習話者をｎ個の話者クラスタにクラスタリングする。そして、各話者クラスタに属する学習話者の音声データを用いた学習によって音響モデルを作成し、

各話者クラスタ別に第１音響モデル格納部７a～第ｎ音響モデル格納部７nに格納するようにしている。

【００６６】その場合、上記学習話者のクラスタリングに際しては、各学習話者間の距離として、上記非線形周波数ワーピング関数f()を用いて、全学習話者の音響モデルに対する学習話者の音響モデルの尤度を最大にするという基準に従って求めた周波数軸の声道長正規化係数αを用いるのである。さらに、発声の仕方や癖によって声門上部の狭めが生じると、広母音の第２フォルマントの存在領域以下の低域のフォルマント周波数が上昇する。そのために、声道長正規化係数αのみでは正しい声道長に写像できないことがある。そこで、上記非線形周波数ワーピング関数f()に、上記低域のフォルマント周波数の上昇に対する補正項としての補正係数βを導入している。

【００６７】そして、標準モデル(全学習話者の音響モデル)と入力モデル(学習話者の音響モデル)の２つの音響モデルにおける対応する状態間の尤度を、標準モデルの出力確率密度関数ｒｉ()に、入力モデルの出力確率密度関数ｂｉ()の平均値ベクトルμｉを非線形周波数ワーピング関数f()で周波数ワープして得られたベクトルを代入したときの値と定義して、上記声道長正規化係数αを上記式(１)によって最尤推定するようにしている。

【００６８】すなわち、本実施の形態によれば、上記学習話者のクラスタリング時に用いる各学習話者間の距離を、生理的な特徴の変動要因である声道長の情報と発声の仕方や癖による影響の補正情報とに基づいて設定することができる。したがって、より少量の発声データから発話者の癖を考慮した正確な各学習話者間の距離に基づいて、学習話者をクラスタリングできるのである。

【００６９】また、実際の学習話者のクラスタリングに際しては、先ずｎ0個の話者クラスタに初期分割し、次にｎ0個の話者クラスタの各境界を中心として上記初期分割された話者クラスタにオーバーラップさせて(ｎ0－１)個の話者クラスタに分割し、合計ｎ＝(２ｎ0－１)個の話者クラスタにクラスタリングするようにしている。したがって、各話者クラスタの境界領域では学習サンプルが希薄だったり段差ができたりしているために上手く学習されず、認識率が劣化するという「harddecision問題」を解消することができるのである。

【００７０】以上のことより、上述のようにしてクラスタリングされた各話者クラスタに属する学習話者の音声データ別に求められた音響モデルを上記第１音響モデル格納部７a～第ｎ音響モデル格納部７nに格納することによって、尤度演算部４は、より入力話者に適合した話者クラスタの音響モデルを適用することができる。したがって、高い認識率を得ることができるのである。

【００７１】尚、上記実施の形態においては、上記話者クラスタ選択部３によって最適な話者クラスタを一つ選択するようにしているが、最適な話者クラスタを含む上位複数の話者クラスタを選択するようにしてもよい。例えば、尤度の上位からｋ個の話者クラスタを選択するとする。そうすると、切換部５によって切り換え選択されたｋ個の音響モデルの夫々に関して、尤度演算部４によって尤度演算が行われて、照合部８にｋ個の尤度系列が送られることになる。したがって、照合部８では、夫々の尤度系列に関して照合処理が行なわれ、最も大きい尤度を呈する単語/単語列が認識結果となるのである。

【００７２】また、音声認識装置のハードウェア規模が大きく、計算量が許すのであれば、話者クラスタ選択部３による話者クラスタ選択を行なわず、尤度演算部において総ての話者クラスタの音響モデルを用いて尤度演算処理を実行するようにしてもよい。この場合、各音響モデルを適用して得られた尤度が最大値を呈する単語/単語列が認識結果となる。

【００７３】＜第２実施の形態＞図４は、本実施の形態の音声認識装置におけるブロック図である。尚、この音声認識装置は、話者正規化方式を用いた音声認識装置である。音声入力部１１,音響分析部１２,尤度演算部１４,照合部１８,辞書格納部１９および出力部２０は、図１に示す上記第１実施の形態における音声入力部１,音響分析部２,尤度演算部４,照合部８,辞書格納部９および出力部１０と同様である。

【００７４】周波数ワープ関数推定部１５は、全話者音響モデル格納部１６に格納された混合数が１の全話者モデル(HMM)を用いて、上記第１実施の形態における話者クラスタ選択方法[ｃ]で述べたように、音響モデルの作成方法における(３)の声道長正規化係数αの導出で説明した手法と同様の手法を用いて、入力音声データから非線形周波数ワーピング関数f()の声道長正規化係数αおよび補正係数βを推定する。そして、推定された声道長正規化係数αおよび補正係数βは、周波数ワープ部１３に送出される。尚、全話者音響モデル格納部１６に格納された全話者モデルは、上記第１実施の形態の音響モデルの作成方法における(１)の全話者モデルの作成で説明した手法と同様の手法で作成される。

【００７５】そうすると、上記周波数ワープ部１３は、上記推定値(α,β)を係数とする非線形周波数ワーピング関数f()を用いて、入力音声の音響パラメータ系列を周波数ワープ(話者正規化)し、周波数ワープ後の音響パラメータ系列を尤度演算部１４に送出するのである。そして、尤度演算部１４では、周波数ワープされた音響パラメータ系列に対して、不特定話者音響モデル格納部１７に格納された不特定話者モデル(HMM)を作用させて、各音韻の状態毎に尤度を算出するのである。

【００７６】ところで、上記周波数ワープ関数推定部１５における上記係数(α,β)の推定方法には、以下に述べる二通りの推定方法がある。

(A) 入力音声データを直接用いる。

(B) 標準話者の音響モデルを入力音声データに話者適

応させた適応音響モデルを用いる。

そして、この二通りの推定方法を、入力音声データの量や質に応じて使い分けるのである。ここで、音声データの質とは尤度の上昇具合であり、周波数ワープ関数推定部１５は、上記二通りの推定方法による尤度の上昇具合を見計らって、上昇の大きい推定方法を採用するのである。長いエンロール期間を許容できる音声認識装置の場合には、このような推定処理も可能となる。尚、長いエンロール期間を許容できない場合には、予め何れかの推定方法に固定しておけばよい。

【００７７】上記推定方法(A)は、入力音声データが多い場合に有効であり、入力音声データから直接求めるために、精密な推定が可能となる。但し、入力音声データが少ない場合には、当該推定をエンロールモードで行う際に入力音声データに無い音素環境における係数(α, β)の推定や平滑化が問題になる。また、推定方法(B)は、入力音声データが少ない場合に有効であり、適応音響モデルの状態毎に補正係数βを制御できるというメリットがある。

【００７８】また、上記推定方法(A),(B)の各々に関して、使用する音響モデルは、全話者モデルの場合と、話者クラスタ別に作成された混合数が１の音響モデルの場合との二通りがある。音声認識装置の記憶容量が少ない場合には前者を採用する。一方、記憶容量が多い場合は音響モデル群を各話者クラスタ別に格納できるので後者を採用する。後者の場合には、入力音声データに基づいて最適な話者クラスタを選択し、この選択話者クラスタに属する音響モデルを使用することになる。すなわち、図４に示す音声認識装置は、全話者モデルを用いた推定方法(A)によって係数(α, β)の推定を行うのである。

【００７９】以上、上記HMMに代表される音響モデルを用いた音声認識装置を例に、本実施の形態を説明したが、標準パターンとして音声波形または音響パラメータ系列を登録しておく音声認識装置に対しても、本実施の形態における話者正規化方法を適用することができる。その場合には、入力音声の音響パラメータ系列で成る特徴パターンと上記標準パターンとのマッチングには、上記HMMの場合の尤度に代ってスペクトル間の距離尺度を用いる。尚、その場合におけるマッチング部による処理手順を図５のフローチャートに示す。以下、図５に従って、標準パターンを登録しておく音声認識装置における上記マッチング部による処理手順について説明する。尚、この場合、係数(α, β)の更新幅と最大値とが予め設定されているものとする。

【００８０】ステップＳ1で、上記特徴パターンと標準パターンの各フレーム間の対応関係（マッチングパスと言う）がＤＰマッチングによって求められる。その場合、上記ＤＰマッチングに際しては、距離尺度としてケプストラム距離等のスペクトル間距離が用いられる。さ

らに、係数(α, β)に初期値が代入される。ステップＳ2で、上記マッチングパスにおける格子点位置が最初の格子点に初期化される。

【００８１】ステップＳ3で、当該格子点に対応するフレームにおける入力音声の音響パラメータ系列に、(α, β)を係数とする非線形周波数ワーピング関数f()を作用させる。こうして、入力音声の当該フレームが係数(α, β)で非線形周波数ワープされる。

【００８２】ステップＳ4で、当該格子点に対応するフレームにおける非線形周波数ワープ後の入力音声の特徴パターンと上記標準パターンとの累積距離が算出される。ステップＳ5で、次の格子点が在るか否かが判別される。その結果、在ればステップＳ6に進み、無ければステップＳ7に進む。ステップＳ6で、上記格子点位置が次の格子点に更新される。そうした後、上記ステップＳ3に戻って、次の格子点での処理に移行する。ステップＳ7で、上記係数(α, β)は上記最大値であるか否かが判別される。その結果、最大値であればステップＳ9に進む一方、そうでなければステップＳ8に進む。ステップＳ8で、係数(α, β)の値が上記更新幅だけ更新される。そうした後に、上記ステップＳ2に戻って、次の係数(α, β)での処理に移行する。このようにして、係数(α, β)の値を上記更新幅だけ順次更新しながら、上記非線形周波数ワープ後の入力音声と標準パターンとにおける上記マッチングパスに沿った累積距離が算出される。そして、上記ステップＳ7において、上記係数(α, β)は上記最大値であると判別されると上記ステップＳ9に進むのである。

【００８３】ステップＳ9で、上記総ての累積距離の算出結果に基づいて、上記非線形周波数ワープ後の入力音声の特徴パターンと標準パターンとの累積距離を最小にする係数(α, β)の値が係数(α, β)の推定値として求められる。ステップＳ10で、上記係数(α, β)の推定値を係数とする上記非線型周波数ワーピング関数f()を用いて、入力音声の音響パラメータ系列を周波数ワープさせる。そして、周波数ワープ後の入力音声の音響パラメータ系列を照合部へ送出して、上記マッチング部による処理を終了するのである。

【００８４】尚、上記マッチング部による処理の説明においては、単純なやり方で係数(α, β)の全部の組合せに関して累積距離を求めているが、山登り法や最急降下法等の高速に収束させる方法を採用しても差し支えない。

【００８５】上述のように、本実施の形態においては、上記周波数ワープ関数推定部１５によって、入力音声の音響パラメータ系列から非線型周波数ワーピング関数f()の係数(α, β)を推定する。そして、周波数ワープ部１３によって、上記推定値(α, β)を係数とする非線形周波数ワーピング関数f()を用いて入力音声の音響パラメータ系列を周波数ワープすることによって、話者正規

21

化するようにしている。

【００８６】その場合、発声の仕方や癖によって声門上部の狭めが生じると、広母音の第２フォルマントの存在領域以下の低域のフォルマント周波数が上昇する。そのために、上記声道長正規化係数αのみを用いた非線形周波数ワーピング関数f()では正しい声道長に写像できないことがある。そこで、上記非線形周波数ワーピング関数f()に、上記低域のフォルマント周波数の上昇に対する補正項としての補正係数βを導入している。

【００８７】そして、全学習話者の音響モデルと入力音声の音響モデルとの２つの音響モデルにおける対応する状態間の尤度を、標準モデルの出力確率密度関数ｒｉ()に、入力モデルの出力確率密度関数ｂｉ()の平均値ベクトルμｉを非線形周波数ワーピング関数f()で周波数ワープして得られたベクトルを代入したときの値と定義して、上記係数(α，β)を上記式(１)によって最尤推定するようにしている。

【００８８】すなわち、本実施の形態によれば、上記話者正規化する際に用いる非線型周波数ワーピング関数f()の係数(α，β)として、生理的な特徴の変動要因である声道長の情報である声道長正規化係数αに対して発声の仕方や癖による影響の補正を行ったものを用いることができる。したがって、発話者の癖を考慮した上記非線型周波数ワーピング関数f()に基づいて、より少量の発声データから、標準話者のスペクトルにより近い周波数特性を有するように入力音声を話者正規化できるのである。

【００８９】＜第３実施の形態＞図６は、本実施の形態の音声認識装置におけるブロック図である。尚、この音声認識装置は、話者適応方式を用いた音声認識装置である。音声入力部２１，音響分析部２２，尤度演算部２７，照合部２９，辞書格納部３０及び出力部３１は、図１に示す上記第１実施の形態における音声入力部１，音響分析部２，尤度演算部４，照合部８，辞書格納部９および出力部１０と同様である。また、周波数ワープ関数推定部２３，全話者音響モデル格納部２４および不特定話者音響モデル格納部２６は、図４に示す上記第２実施の形態における周波数ワープ関数推定部１５，全話者音響モデル格納部１６および不特定話者音響モデル格納部１７と同様である。

【００９０】すなわち、上記周波数ワープ関数推定部２３は、上記第２実施の形態の場合と同様にして、全話者モデルを用いて、入力音声データから非線形周波数ワーピング関数f()の係数(α，β)を推定する。そして、周波数ワープ部２５によって、この推定された係数(α，β)の逆数を係数とする非線形周波数ワーピング関数f()を用いて、不特定話者音響モデル格納部２６に格納された不特定話者モデルを周波数ワープする。こうして周波数ワープされた不特定話者音響モデルを、話者適応モデル(HMM)として話者適応音響モデル格納部２８に格納する。そうすると、尤度演算部２７は、音響分析部２２からの入力音声の音響パラメータ系列に対して、話者適応音響モデル格納部２８に格納された話者適応モデルを作用させて、上述した尤度演算処理を行なうのである。

22

【００９１】このように、本実施の形態においては、上記周波数ワープ関数推定部２３によって、入力音声の音響パラメータ系列から非線型周波数ワーピング関数f()の係数(α，β)を推定する。そして、周波数ワープ部２５によって、上記推定値(α，β)の逆数を係数とする非線形周波数ワーピング関数f()を用いて不特定話者モデルを周波数ワープすることによって、不特定話者モデルを話者適応させるようにしている。

【００９２】その場合、発声の仕方や癖によって声門上部の狭めが生じると、広母音の第２フォルマントの存在領域以下の低域のフォルマント周波数が上昇する。そのために、上記声道長正規化係数αのみを用いた非線形周波数ワーピング関数f()では正しい声道長に写像できないことがある。そこで、上記非線形周波数ワーピング関数f()に、上記低域のフォルマント周波数の上昇に対する補正項としての補正係数βを導入している。

【００９３】そして、全学習話者の音響モデルと入力音声の音響モデルとの２つの音響モデルにおける対応する状態間の尤度を、標準モデルの出力確率密度関数ｒｉ()に、入力モデルの出力確率密度関数ｂｉ()の平均値ベクトルμｉを非線形周波数ワーピング関数f()で周波数ワープして得られたベクトルを代入したときの値と定義して、上記係数(α，β)を上記式(１)によって最尤推定するようにしている。

【００９４】すなわち、本実施の形態によれば、上記不特定話者モデルを話者適応する際に用いる非線型周波数ワーピング関数f()の係数(α，β)として、生理的な特徴の変動要因である声道長の情報である声道長正規化係数αに対して発声の仕方や癖による影響の補正を行ったものを用いることができる。したがって、発話者の癖を考慮した上記非線型周波数ワーピング関数f()に基づいて、より少量の発声データから入力話者のスペクトルにより近い周波数特性を有するように不特定話者モデルを話者適応できるのである。

【００９５】尚、本実施の形態における上記話者適応音響モデル格納部２８に格納する話者適応モデルの与え方には、上述の与え方の以外に、話者クラスタを用いる方法を採用してもよい。そして、この二通りの与え方を、音声認識装置の規模や入力音声データの量や質に応じて使い分けるのである。ここで、音声データの質とは尤度の上昇具合であり、周波数ワープ関数推定部２３は、上記二通りの与え方による尤度の上昇具合を見計らって、上昇の大きい推定方法を採用するのである。長いエンロール期間が許容できる音声認識装置の場合には、このような推定処理も可能となる。尚、上記話者クラスタを用いる方法においては、教師語彙を与える上記第１実施の

形態における選択法[ａ]に基づいて話者クラスタを選択する。そして、選択された話者クラスタの音響モデルを話者適応モデルとして話者適応音響モデル格納部２８に格納するのである。

【００９６】また、上述した二つの与え方の何れかによって得られた話者適応モデルを初期モデルとして、上記ＭＬＬＲ方やＶＦＳ法等の既存の話者適応技術を用いて話者適応を行って新たに話者適応モデルを生成し、これを尤度演算部で用いるようにしても差し支えない。尚、その際における話者クラスタのクラスタ数や補正係数βの使い方(固定範囲)やクラスタ選択の方法等は、適用する音声認識装置あるいは用いる音響モデルの規模やエンロールモードの有無等に依存するため、それらの項目については、本実施の形態においては規定しない。例えば、コンパクトな音声認識装置を望む場合には、上記話者クラスタの数は減ることになる。また、音響モデルの規模が小さい場合には、補正係数βは状態毎に固定すればよい。エンロールモードが許容できない場合には、ＧＭＭを用いたクラスタ選択が行われることになる。

【００９７】＜第４実施の形態＞図７は、本実施の形態のテキスト音声合成装置におけるブロック図である。なお、このテキスト音声合成装置は、声質変換方式を用いたテキスト音声合成装置である。テキスト解析部４１は、単語とそのアクセント型とが格納されたアクセント辞書４２を用い、入力テキストに対して形態素解析および係り受け解析を行って音素文字列とアクセント情報とを生成して韻律生成部４３に送出する。韻律生成部４３は、韻律制御テーブル４４を参照して、継続時間長やピッチやパワーの韻律情報を生成して、音素文字列と共に音声素片選択部４５に送出する。そうすると、音声素片選択部４５は、音声素片辞書４６から音素環境や韻律環境に最適な音声素片を選択し、音声素片情報を生成する。そして、この生成された音声素片情報を周波数ワープ部４８に出力する一方、上記韻律情報を音声素片合成部４７に出力する。

【００９８】一方、周波数ワープ関数推定部４９は、声質変換のターゲット話者の入力音声波形を基に、第２，第３実施の形態の場合と同様にして、上記非線形周波数ワーピング関数ｆ()の係数(α，β)を推定する。そうすると、周波数ワープ部４８は、この推定された係数(α，β)の逆数を係数とする非線形周波数ワーピング関数ｆ()を用いて上記音声素片情報である音響パラメータ系列を周波数ワープし、周波数ワープ後の音声素片情報を音声素片合成部４７に送出する。最後に、音声素片合成部４７は、周波数ワープ部４８からの周波数ワープ後の音声素片情報(音声素片の音響パラメータ系列)と音声素片選択部４５からの韻律情報とを用いて、音声波形を生成しスピーカ５０から音声出力する。

【００９９】上述のように、本実施の形態においては、テキスト音声合成を行うに際して、上記周波数ワープ関数推定部４９によって、声質変換のターゲット話者における入力音声の音響パラメータ系列から非線型周波数ワーピング関数ｆ()の係数(α，β)を推定する。そして、周波数ワープ部４８によって、上記推定値(α，β)を係数とする非線形周波数ワーピング関数ｆ()を用いて、テキストに基づいて選択された音声素片の音響パラメータ系列を周波数ワープすることによって、声質変換を行うようにしている。

【０１００】その場合、上記係数(α，β)を推定に際しては、発声の仕方や癖によって声門上部の狭めが生じると、広母音の第２フォルマントの存在領域以下の低域のフォルマント周波数が上昇する。そのために、上記非線形周波数ワーピング関数ｆ()に、上記低域のフォルマント周波数の上昇に対する補正項としての補正係数βを導入している。

【０１０１】そして、全学習話者の音響モデルと入力音声の音響モデルとの２つの音響モデルにおける対応する状態間の尤度を、標準モデルの出力確率密度関数ｒｉ()に、入力モデルの出力確率密度関数ｂｉ()の平均値ベクトルμｉを非線形周波数ワーピング関数ｆ()で周波数ワープして得られたベクトルを代入したときの値と定義して、上記声道長正規化係数αを、上記式(１)によって最尤推定するようにしている。

【０１０２】すなわち、本実施の形態によれば、上記声質変換を行う際に用いる非線型周波数ワーピング関数ｆ()の係数(α，β)として、生理的な特徴の変動要因である声道長の情報である声道長正規化係数αに対して発声の仕方や癖による影響の補正を行ったものを用いることができる。したがって、発話者の癖を考慮した上記非線型周波数ワーピング関数ｆ()に基づいて、より少量の発声データから、ターゲット話者のスペクトルにより近い周波数特性を有するように音声素片情報を声質変換できるのである。

【０１０３】本実施の形態はスペクトル包絡の変換であり、声質の適応におおいに効果がある。しかしながら、話者間の声の特徴差は声質だけでなはく韻律が大きく寄与する。したがって、本実施の形態に対して韻律の適応技術を併用しても構わない。

【０１０４】尚、上述した各実施の形態においては、上記声道長正規化係数αと補正係数βとで成る話者特徴を用いてクラスタリングされた音響モデルを搭載した音声認識装置、上記声道長正規化係数αと補正係数βとで成る話者特徴を用いて話者正規化あるいは話者適応を行う音声認識装置、および、上記声道長正規化係数αと補正係数βとで成る話者特徴を用いて声質変換を行う音声合成装置について説明している。しかしながら、この発明は、上記声道長正規化係数αと補正係数βとを話者特徴として抽出する話者特徴抽出装置にも適用されるものである。

【０１０５】ところで、その場合の話者特徴抽出装置に

おける上記伸縮係数取得手段としての機能は、プログラム記録媒体に記録された話者特徴抽出処理プログラムによって実現される。上記プログラム記録媒体は、ＲＯＭ(リード･オンリ･メモリ)でなるプログラムメディアである。あるいは、外部補助記憶装置に装着されて読み出されるプログラムメディアであってもよい。尚、何れの場合においても、上記プログラムメディアから話者特徴抽出処理プログラムを読み出すプログラム読み出し手段は、上記プログラムメディアに直接アクセスして読み出す構成を有していてもよいし、ＲＡＭ(ランダム･アクセス･メモリ)に設けられたプログラム記憶エリア(図示せず)にダウンロードして、上記プログラム記憶エリアにアクセスして読み出す構成を有していてもよい。尚、上記プログラムメディアからＲＡＭの上記プログラム記憶エリアにダウンロードするためのダウンロードプログラムは、予め本体装置に格納されているものとする。

【０１０６】ここで、上記プログラムメディアとは、本体側と分離可能に構成され、磁気テープやカセットテープ等のテープ系、フロッピー（登録商標）ディスク,ハードディスク等の磁気ディスクやＣＤ(コンパクトディスク)－ＲＯＭ,ＭＯ(光磁気)ディスク,ＭＤ(ミニディスク),ＤＶＤ(ディジタルビデオディスク)等の光ディスクのディスク系、ＩＣ(集積回路)カードや光カード等のカード系、マスクＲＯＭ,ＥＰＲＯＭ（紫外線消去型ＲＯＭ),ＥＥＰＲＯＭ(電気的消去型ＲＯＭ),フラッシュＲＯＭ等の半導体メモリ系を含めた、固定的にプログラムを坦持する媒体である。

【０１０７】また、上記各実施の形態における音声認識装置,音声合成装置および話者特徴抽出装置は、モデムを備えてインターネットを含む通信ネットワークと接続可能な構成を有していれば、上記プログラムメディアは、通信ネットワークからのダウンロード等によって流動的にプログラムを坦持する媒体であっても差し支えない。尚、その場合における上記通信ネットワークからダウンロードするためのダウンロードプログラムは、予め本体装置に格納されているものとする。または、別の記録媒体からインストールされるものとする。

【０１０８】尚、上記記録媒体に記録されるものはプログラムのみに限定されるものではなく、データも記録することが可能である。

【０１０９】

【発明の効果】以上より明らかなように、第１の発明の話者特徴抽出装置は、伸縮係数取得手段によって、音声のスペクトルにおける広母音の第２フォルマントの存在領域以下の低い周波数領域で伸縮係数αに補正係数βを乗じて補正を行った非線形周波数ワーピング関数を用いて、標準話者の音声パターンに対して入力話者の音声パターンの尤度を最大にするという基準に従って上記伸縮係数αを求め、求めた伸縮係数αを話者特徴とするので、生理的な特徴の変動要因である声道長の情報に対して発声の仕方や癖による影響の補正を行って、より話者に適合した特徴を抽出することができる。さらに、上記発声の仕方や癖を表す発声データを必要とはせず、より少量の発声データから精度良く話者特徴を抽出できる。

【０１１０】また、上記第１の発明の話者特徴抽出装置は、上記伸縮係数取得手段を、上記非線形周波数ワーピング関数を用いて、標準話者の音声パターンに対して入力話者の音声パターンの尤度を最大にするという基準に従って上記補正係数βをも求めるように成せば、さらに話者に適合した特徴を抽出することができる。

【０１１１】また、第２の発明の音声認識装置は、音響モデルを、上記伸縮係数αに基づいて学習話者をクラスタリングして得られた各話者クラスタ別に格納する際に、上記伸縮係数αを、上記非線形周波数ワーピング関数を用いた最尤推定によって求めるので、生理的な特徴の変動要因である声道長の情報に対して発声の仕方や癖による影響の補正を行って、より学習話者の音声パターンに適合した学習話者間の距離を用いて上記クラスタリングを行うことができる。したがって、この発明によれば高い認識率を得ることができる。さらに、上記発声の仕方や癖を表す発声データを必要とはせず、より少量の発声データから学習話者間の距離を得ることができるのである。

【０１１２】また、上記第２の発明の音声認識装置は、上記学習話者のクラスタリングを上記伸縮係数αと補正係数βとの２次元平面に対して行うようになっており、上記補正係数βを上記非線形周波数ワーピング関数を用いた最尤推定によって求めれば、さらに学習話者の音声パターンに適合した距離を用いてクラスタリングを行うことができる。

【０１１３】また、上記第２の発明の音声認識装置は、上記話者クラスタを、所定のクラスタ数の初期クラスタと、上記各初期クラスタの境界を含んで上記各初期クラスタにオーバーラップするオーバーラップクラスタとで構成すれば、学習サンプルが希薄だったり段差ができ易い上記各初期クラスタの境界領域を、何れかのオーバーラップクラスタに属させることができる。したがって、上記各初期クラスタの境界領域において認識率が劣化するという「hard decision問題」を解消できる。

【０１１４】また、第３の発明の音声認識装置は、正規化手段を、上記非線形周波数ワーピング関数を用いて伸縮係数αと補正係数βとを最尤推定する周波数ワーピング関数推定手段と、上記推定されたαとβとを係数とする上記非線形周波数ワーピング関数を用いて上記入力話者の音声スペクトルの周波数軸を伸縮する周波数ワープ手段で構成したので、より標準話者の音声スペクトルに近づくように話者を正規化することができる。したがって、この発明によれば高い認識率を得ることができる。さらに、上記発声の仕方や癖を表す発声データを必要とはせず、より少量の発声データに基づいて話者正規化を

行うことができる。

【０１１５】また、第４の発明の音声認識装置は、話者適応手段を、上記非線形周波数ワーピング関数を用いて伸縮係数αと補正係数βとを最尤推定する周波数ワーピング関数推定手段と、上記推定されたαの逆数とβの逆数とを係数とする上記非線形周波数ワーピング関数を用いて音響モデルの周波数軸を伸縮する周波数ワープ手段で構成したので、より入力話者の音声スペクトルに近づくように話者適応を行うことができる。したがって、この発明によれば、高い認識率を得ることができる。さらに、上記発声の仕方や癖を表す発声データを必要とはせず、より少量の発声データに基づいて話者適応を行うことができる。

【０１１６】また、上記第３の発明あるいは第４の発明の音声認識装置は、上記周波数ワーピング関数推定手段を、上記入力話者の音声パターンの代わりに、標準話者の音響モデルを上記入力話者の音声パターンに話者適応させた適応音響モデルを用いるように成せば、入力話者の音声パターン数が少ない場合でも対処することができる。さらに、上記適応音響モデルの状態毎に補正係数βを制御して、話者の発声の仕方や癖による入力音声パターンのずれを木目細かく補正することが可能になる。

【０１１７】また、上記第２の発明乃至第４の発明の何れか一つの発明の音声認識装置は、上記補正係数βをサブワード単位に求め、上記サブワード毎に決定すれば、上記補正係数βを上記サブワード単位で変更することができ、話者の発声の仕方や癖による入力音声パターンのずれを木目細かく補正することができる。

【０１１８】また、第５の発明の音声合成装置は、声質変換手段を、上記非線形周波数ワーピング関数を用いて伸縮係数αと補正係数βとを最尤推定する周波数ワーピング関数推定手段と、上記推定されたαの逆数とβの逆数とを係数とする上記非線形周波数ワーピング関数を用いて標準話者の音声素片の周波数軸を伸縮する周波数ワープ手段で構成したので、より発話者の声質に適合するように合成音声の声質を変換することができる。さらに、上記発声の仕方や癖を表す発声データを必要とはせず、より少量の発声データに基づいて声質変換を行うことができる。

【０１１９】また、上記第５の発明の音声合成装置は、上記周波数ワーピング関数推定手段をサブワード単位に上記補正係数βを求め、上記サブワード毎に上記補正係数βを推定するように成せば、上記補正係数βを上記サブワード単位に変更することができ、発話者の発声の仕方や癖による入力音声パターンのずれを木目細かく補正することができる。

【０１２０】また、第６の発明の話者特徴抽出方法は、上記非線形周波数ワーピング関数を用いて、最尤度推定によって上記伸縮係数αを求めて話者特徴とするので、生理的な特徴の変動要因である声道長の情報に対して発声の仕方や癖による影響の補正を行って、より話者に適合した特徴を抽出することができる。さらに、上記発声の仕方や癖を表す発声データを必要とはせず、より少量の発声データから精度良く話者特徴を抽出できる。

【０１２１】また、第７の発明のプログラム記録媒体は、コンピュータを、上記第１の発明における上記伸縮係数取得手段として機能させる話者特徴抽出処理プログラムが記録されているので、上記第１の発明の場合と同様に、より話者に適合した特徴を抽出することができる。さらに、より少量の発声データから良質の話者特徴を抽出できる。

【図面の簡単な説明】

【図１】　この発明の話者クラスタリング方式を用いた音声認識装置におけるブロック図である。

【図２】　非線形周波数ワーピング関数の一例を示す図である。

【図３】　初期分割数が５である場合のクラスタリング例を示す図である。

【図４】　図１とは異なる話者正規化方式を用いた音声認識装置のブロック図である。

【図５】　標準パターンを用いる音声認識装置に図４と同様の話者正規化方式を適用した際におけるマッチング部による処理手順のフローチャートである。

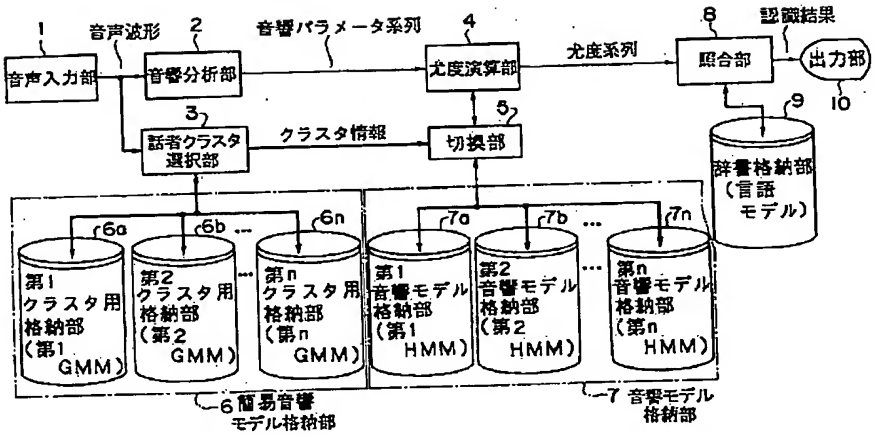
【図６】　図１および図４とは異なる話者適応方式を用いた音声認識装置におけるブロック図である。

【図７】　この発明の声質変換方式を用いた音声合成装置におけるブロック図である。

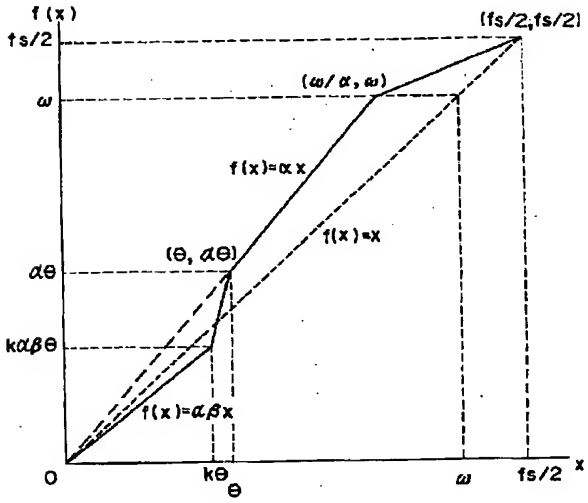
【符号の説明】

１，１１，２１…音声入力部、
２，１２，２２…音響分析部、
３…話者クラスタ選択部、
４，１４，２７…尤度演算部、
５…切換部、
６…簡易音響モデル格納部、
７…音響モデル格納部、
８，１８，２９…照合部、
９，１９，３０…辞書格納部、
１０，２０，３１…出力部、
１３，２５，４８…周波数ワープ部、
１５，２３，４９…周波数ワープ関数推定部、
１６，２４…全話者音響モデル格納部、
１７，２６…不特定話者音響モデル格納部、
２８…話者適応音響モデル格納部、
４１…テキスト解析部、
４３…韻律生成部、
４５…音声素片選択部、
４７…音声素片合成部、
５０…スピーカ。

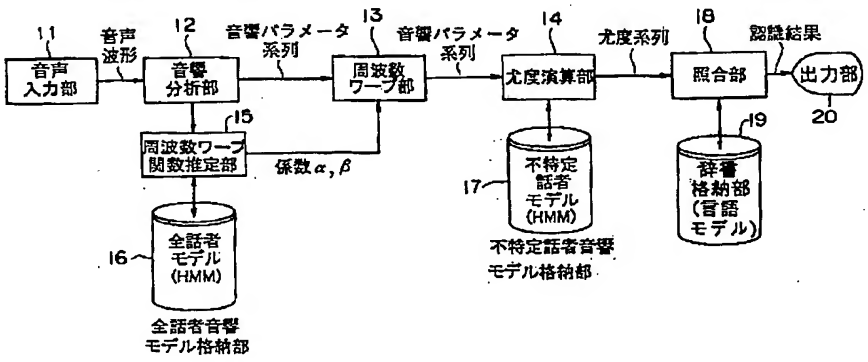
【図１】


1
音声入力部
音声波形
2
音響分析部
音響パラメータ系列
4
尤度演算部
尤度系列
8
照合部
認識結果
出力部
10
3
話者クラスタ選択部
クラスタ情報
5
切換部
9
辞書格納部（言語モデル）
6a
6b
6n
第1クラスタ用格納部（第1 GMM）
第2クラスタ用格納部（第2 GMM）
第nクラスタ用格納部（第n GMM）
7a
7b
7n
第1音響モデル格納部（第1 HMM）
第2音響モデル格納部（第2 HMM）
第n音響モデル格納部（第n HMM）
6 簡易音響モデル格納部
7 音響モデル格納部

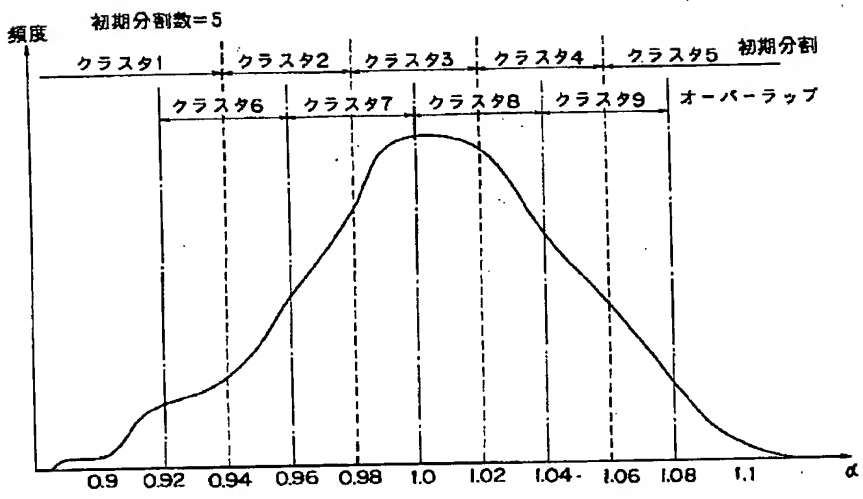
【図２】


f(x)
fs/2
(fs/2,fs/2)
ω
(ω/α, ω)
f(x)=αx
f(x)=x
(θ, αθ)
αθ
kαβθ
f(x)=αβx
0
kθ
θ
ω
fs/2
x

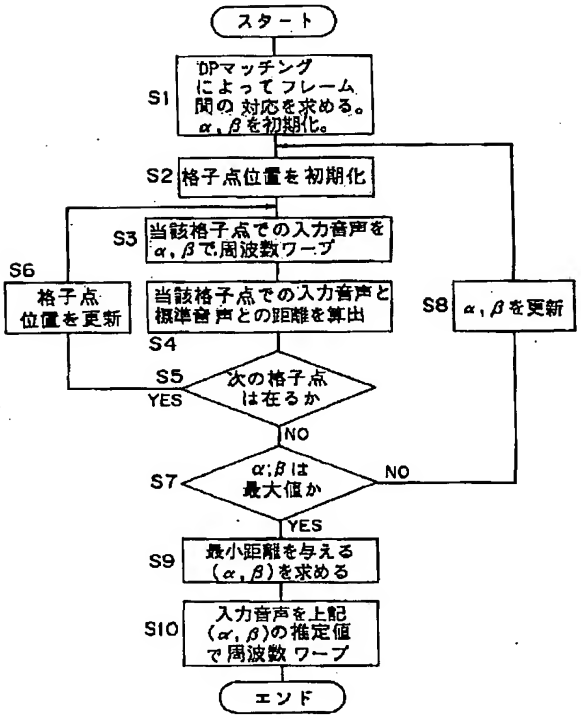
【図４】


11
音声入力部
音声波形
12
音響分析部
音響パラメータ系列
13
周波数ワープ部
音響パラメータ系列
14
尤度演算部
尤度系列
18
照合部
認識結果
出力部
20
15
周波数ワープ関数推定部
係数α, β
16
全話者モデル(HMM)
全話者音響モデル格納部
17
不特定話者モデル(HMM)
不特定話者音響モデル格納部
19
辞書格納部（言語モデル）

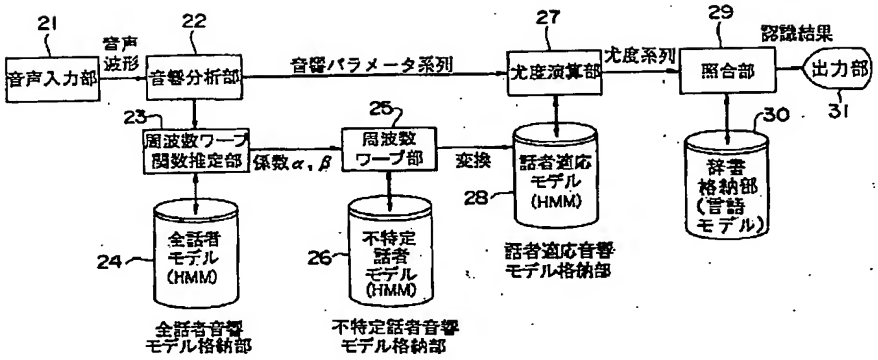
【図３】


頻度
初期分割数＝5
クラスタ1
クラスタ2
クラスタ3
クラスタ4
クラスタ5
初期分割
クラスタ6
クラスタ7
クラスタ8
クラスタ9
オーバーラップ
0.9
0.92
0.94
0.96
0.98
1.0
1.02
1.04
1.06
1.08
1.1
α

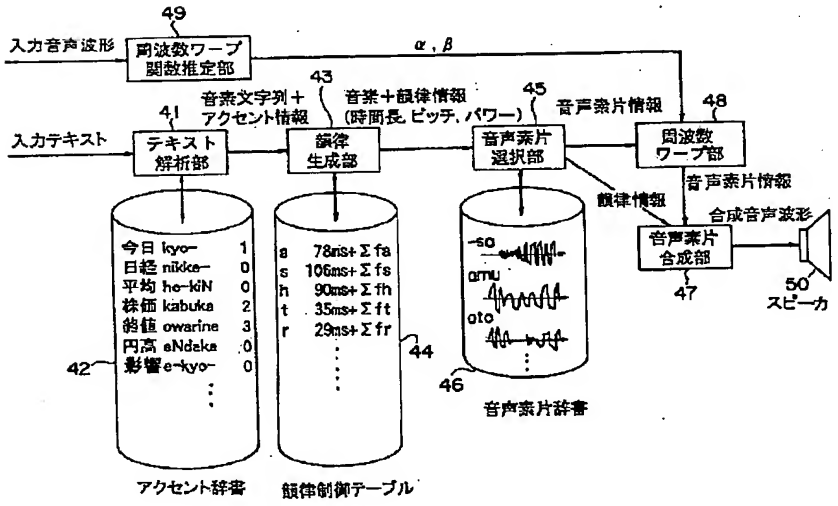
【図５】


スタート
S1
DPマッチング によってフレーム 間の対応を求める。 α，βを初期化。
S2
格子点位置を初期化
S3
当該格子点での入力音声を α，βで周波数ワープ
S4
当該格子点での入力音声と 標準音声との距離を算出
S5
次の格子点 は在るか
YES
NO
S6
格子点 位置を更新
S7
α，βは 最大値か
NO
YES
S8
α，βを更新
S9
最小距離を与える (α，β)を求める
S10
入力音声を上記 (α，β)の推定値 で周波数ワープ
エンド

【図６】



【図７】



フロントページの続き

(51) Int. Cl.[7]	識別記号	F I	テーマコード（参考）
		G10L 9/10	301A